Agenda: Finish term-frequency weighting; tf-idf weighting.

I. Reminder: term-frequency weighting

Define \( \text{freq}_{i,j} \) (term-document frequency) as the number of times term \( w_i \) occurs in document \( d_j \).

We then set the document vector \( \vec{d}_j \) for document \( d_j \) as follows:

\[
\vec{d}_j = \left( \frac{\text{freq}_{1,j}}{L_j}, \frac{\text{freq}_{2,j}}{L_j}, \ldots, \frac{\text{freq}_{m,j}}{L_j} \right)
\]

where \( L_j = \sqrt{\sum_{i=1}^{m} \text{freq}_{i,j}^2} \) is the length-normalization factor.

II. Example corpus and query from last time

Vocabulary: \( w_1 \): cats; \( w_2 \): dogs; \( w_3 \): news

\( d_4 \): “cats news”
\( d_5 \): “cats news cats news”
\( d_6 \): “cats dogs news news dogs”

\( q \): “cats dogs”

III. Inverse document frequency

We define \( \text{IDF}_i \), the inverse document frequency of term \( w_i \), as

\[
\text{IDF}_i = \frac{n}{\text{docfreq}_i},
\]

where \( \text{docfreq}_i \) is the number of documents in the \( n \)-document corpus that contain \( w_i \).

For example, suppose we have \( w_1 = “\text{Bill}” \), \( w_2 = “\text{the}” \), and \( w_3 = “\text{I}” \), and we have in our corpus just the following two documents, where we’ve highlighted occurrences of \( w_1 \), \( w_2 \), and \( w_3 \):

\( d_1 \): Bill Gates of Microsoft spoke at yesterday’s convention. We were kind of surprised at some of the predictions he made, but later on some other presentations clarified the situation. After all, the industry’s followed these trends so far.

\( d_2 \): My friend Bill says weird versions of common proverbs. Just the other day, he said “Gates make for good neighbors.” I also heard him say, “Microsoft wasn’t built in a day”, which is true, I have to admit.

We have \( \text{IDF}_1 = 1 \) and \( \text{IDF}_3 = 2 \). Note that we would get the same IDF for \( w_3 = “\text{I}” \) whether “I” occurred once in \( d_2 \) or 40 times.

IV. Tf-idf weighting:

This alternative to term-frequency weighting converts a document \( d_j \) to the vector

\[
\vec{d}_j = \left( \frac{\text{freq}_{1,j} \times \text{IDF}_1}{L_j}, \frac{\text{freq}_{2,j} \times \text{IDF}_2}{L_j}, \ldots, \frac{\text{freq}_{m,j} \times \text{IDF}_m}{L_j} \right)
\]

where \( L_j = \sqrt{\sum_{i=1}^{m} \left( \text{freq}_{i,j} \times \text{IDF}_i \right)^2} \).